physical sciences forum

MDPI

Proceeding Paper

Value of Information in the Binary Case and Confusion Matrix [†]

Roman Belavkin [1,*], Panos Pardalos [2] and Jose Principe [3]

[1] Faculty of Science and Technology, Middlesex University, London NW4 4BT, UK
[2] Department of Industrial & Systems Engineering, University of Florida, P.O. Box 116595, Gainesville, FL 32611-6595, USA
[3] Department of Electrical & Computer Engineering, University of Florida, P.O. Box 116130, Gainesville, FL 32611-6130, USA
[*] Correspondence: r.belavkin@mdx.ac.uk; Tel.: +44-208-411-6263

[†] Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

Abstract: The simplest Bayesian system used to illustrate ideas of probability theory is a coin and a boolean utility function. To illustrate ideas of hypothesis testing, estimation or optimal control, one needs to use at least two coins and a confusion matrix accounting for the utilities of four possible outcomes. Here we use such a system to illustrate the main ideas of Stratonovich's value of information (VoI) theory in the context of a financial time-series forecast. We demonstrate how VoI can provide a theoretical upper bound on the accuracy of the forecasts facilitating the analysis and optimization of models.

Keywords: value of information; Shannon's information; confusion matrix; time-series forecast

check for updates

Citation: Belavkin, R.; Pardalos, P.; Principe, J. Value of Information in the Binary Case and Confusion Matrix. *Phys. Sci. Forum* **2022**, *5*, 8. https://doi.org/10.3390/psf2022005008

Academic Editors: Frédéric Barbaresco, Ali Mohammad-Djafari, Frank Nielsen and Martino Trassinelli

Published: 2 November 2022

1. Introduction

The concept of value of information has different definitions in the literature [1,2]. Here we follow the works of Ruslan Stratonovich and his colleagues, who were inspired by Shannon's work on rate distortion [3] and made a number of important developments in the 1960s [2]. These mainly theoretical results are gaining new interest thanks to the advancements in data science and machine learning and the need for a deeper understanding of the role of information in learning. We shall review the value of information theory in the context of optimal estimation and hypothesis testing, although the context of optimal control is also relevant.

Consider a probability space $(\Omega, P, \mathcal{A})$ and a random variable $x : \Omega \to X$ (a measurable function). The optimal estimation of $x \in X$ is the problem of finding an element $y \in Y$ maximizing the expected value of some *utility* function $u : X \times Y \to \mathbb{R}$ (or minimizing for cost $-u$). The optimal value is

$$U(0) := \sup_{y \in Y} \mathbb{E}_{P(x)}\{u(x,y)\}\,,$$

where zero designates the fact that no information about the specific value of $x \in X$ is given, only the prior distribution $P(x)$. At the other extreme, let $z \in Z$ be another random variable that communicates full information about each realization of x. This entails that there is an invertible function $z = f(x)$ such that $x = f^{-1}(z)$ is determined uniquely by the 'message' $z \in Z$. The corresponding optimal value is

$$U(\infty) := \mathbb{E}_{P(x)}\{\sup_{y(z)} u(x, y(z))\}\,,$$

where an optimal y is found for each z (i.e., optimization over all mappings $y : Z \to Y$). In the context of estimation, variable x is the *response* (i.e., the variable of interest) and z is the *predictor*. The mapping $y(z)$ represents a model with output $y \in Y$.

Let $I \in [0, \infty]$ be the intermediate amounts of information, and let $U(I) \in [U(0), U(\infty)]$ be the corresponding optimal values. The *value of information* is the difference [4]:

$$V(I) := U(I) - U(0) .$$

There are, however, different ways in which the information amount I and the quantity $U(I)$ can be defined, leading to different types of the value function $V(I)$. For example, consider a mapping $f : X \to Z$ with a constraint $|Z| \leq e^I < |X|$ on the cardinality of its image. The mapping f partitions its domain into a finite number of subsets $f^{-1}(z) = \{x \in X : f(x) = z\}$. Then, given a specific partition $z(x)$, one can find optimal $y(z)$ maximizing the conditional expected utility $\mathbb{E}_{P(x|z)}\{u(x,y) \mid z\}$ for each subset $f^{-1}(z) \ni x$. This optimization should be repeated for different partitions $z(x)$, and the optimal value $U(I)$ is defined over all partitions $z(x)$, satisfying the cardinality constraint $\ln |Z| \leq I$:

$$U(I) := \sup_{z(x)} \left[\mathbb{E}_{P(z)} \left\{ \sup_{y(z)} \mathbb{E}_{P(x|z)}\{u(x,y) \mid z\} \right\} : \ln |Z| \leq I \right] \tag{1}$$

Here, $P(z) = P\{x \in f^{-1}(z)\}$. The quantity $I = \ln |Z|$ is called *Hartley's information*, and the difference $V(I) = U(I) - U(0)$ in this case is the value of Hartley's information. One can relax the cardinality constraint and replace it with the constraint on entropy $H(Z) \leq I$, where $H(Z) = -\mathbb{E}_{P(z)}\{\ln P(z)\} \leq \ln |Z|$. In this case, $V(I)$ is called the value of *Boltzmann's information* [4].

One can see from Equation (1) that the computation of the value of Hartley's or Boltzmann's information is quite demanding and may involve a procedure such as the k-means clustering algorithm or training a multilayer neural network. Thus, using these values of information is not practical due to high computational costs. The main result of Stratonovich's theory [4] is that the upper bound on Hartley's or Boltzmann's values of information is given by the value of Shannon's information, and that asymptotically all these values are equivalent (Theorems 11.1 and 11.2 in [4]). The value of Shannon's information is much easier to compute.

Recall the definition of Shannon's mutual information [3]:

$$I(X,Y) := \mathbb{E}_{W(x,y)} \left\{ \ln \frac{P(x \mid y)}{P(x)} \right\} = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) ,$$

where $W(x,y) = P(x \mid y)Q(y)$ is the joint probability distribution on $X \times Y$, and $H(X \mid Y)$ is the conditional entropy. Under broad assumptions on the reference measures (see Theorem 1.16 in [4]), the following inequalities are valid:

$$0 \leq I(X,Y) \leq \min\{H(X), H(Y)\} \leq \min\{\ln |X|, \ln |Y|\} .$$

The value of Shannon's information is defined using the quantity:

$$U(I) := \sup_{P(y|x)} [\mathbb{E}_W\{u(x,y)\} : I(X,Y) \leq I] \tag{2}$$

The optimization above is over all conditional probabilities $P(y \mid x)$ (or joint measures $W(x,y) = P(y \mid x)P(x)$) satisfying the information constraint $I(X,Y) \leq I$. Contrast this with $U(I)$ for Hartley's or Boltzmann's information (1), where optimization is over the mappings $y(x) = y \circ z(x)$. As was pointed out in [5], the relation between functions (1) and (2) is similar to that between optimal transport problems in the Monge and Kantorovich formulations. Joint distributions optimal in the sense of (2) are found using the standard method of Lagrange multipliers (e.g., see [4,6]):

$$W(x,y;\beta) = P(x)Q(y)e^{\beta\, u(x,y) - \gamma(\beta,x)} , \tag{3}$$

where parameter β^{-1}, called *temperature*, is the Lagrange multiplier associated with the constraint $I(X,Y) \leq I$. Distributions P and Q are the marginals of W, and function $\gamma(\beta,x)$ is defined by normalization $\sum_{x,y} W(x,y;\beta) = 1$. In fact, taking partial traces of solution (3) gives two equations:

$$\sum_x W(x,y) = Q(y) \quad \Longrightarrow \quad \sum_x e^{\beta\, u(x,y)-\gamma(\beta,x)}\, P(x) = 1 \tag{4}$$

$$\sum_y W(x,y) = P(x) \quad \Longrightarrow \quad \sum_y e^{\beta\, u(x,y)}\, Q(y) = e^{\gamma(\beta,x)} \tag{5}$$

Equation (5) defines function $\gamma(\beta,x) = \ln \sum_y e^{\beta\, u(x,y)}\, Q(y)$. If the linear transformation $T(\cdot) = \sum_x e^{\beta\, u(x,y)}(\cdot)$ has an inverse, then from Equation (4) one obtains $e^{-\gamma(\beta,x)} P(x) = T^{-1}(1)$ or

$$\gamma(\beta,x) = -\ln \sum_y b(x,y) + \ln P(x) = \gamma_0(\beta,x) - h(x)\,,$$

where $\gamma_0(\beta,x) := -\ln \sum_y b(x,y)$, $b(x,y)$ is the kernel of the inverse transformation T^{-1}, and $h(x) = -\ln P(x)$ is random entropy or *surprise*. Integrating the above with respect to measure $P(x)$ we obtain

$$\Gamma(\beta) := \sum_x \gamma(\beta,x)\, P(x) = \Gamma_0(\beta) - H(X)\,,$$

where $\Gamma_0(\beta) := \sum_x \gamma_0(\beta,x)\, P(x)$. Function $\Gamma(\beta)$ is the *cumulant generating function* of optimal distribution (3). Indeed, the expected utility and Shannon's information for this distribution are

$$U(\beta) = \Gamma'(\beta) = \Gamma_0'(\beta)\,, \qquad I(\beta) = \beta\, \Gamma'(\beta) - \Gamma(\beta) = H(X) - [\Gamma_0(\beta) - \beta\, \Gamma_0'(\beta)]\,.$$

The first formula can be obtained directly by differentiating $\Gamma(\beta)$, and the second by substitution of (3) into the formula for Shannon's mutual information. Function $\Gamma_0(\beta) - \beta\, \Gamma_0'(\beta)$ is clearly the conditional entropy $H(X \mid Y)$ because $I(X,Y) = H(X) - H(X \mid Y)$.

Note that information is the Legendre–Fenchel transform $I(U) = \sup\{\beta\, U - \Gamma(\beta)\}$ of convex function $\Gamma(\beta)$ (indeed, $U = \Gamma'(\beta)$). The inverse of $I(U)$ is the optimal value $U(I)$ from Equation (2) defining the value of Shannon's information, and it is the Legendre–Fenchel transform $U(I) = \inf\{\beta^{-1}\, I - F(\beta^{-1})\}$ of concave function $F(\beta^{-1}) = -\beta^{-1}\Gamma(\beta)$, which is called *free energy*.

The general strategy for computing the value of Shannon's information is to derive the expressions for $U(\beta)$ and $I(\beta)$ from function $\Gamma_0(\beta)$ (alternatively, one can obtain $U(\beta^{-1})$ and $I(\beta^{-1})$ from free energy $F_0(\beta^{-1}) = -\beta^{-1}\Gamma_0(\beta)$). Then the dependency $U(I)$ is obtained either parametrically or by excluding β. Let us now apply this to the simplest 2 × 2 case.

2. Value of Shannon's Information for the 2×2 System

Let $X \times Y = \{x_1, x_2\} \times \{y_1, y_2\}$, and let $u : X \times Y \to \mathbb{R}$ be the utility function, which we can represent by a 2×2 matrix:

$$\|u(x,y)\| = \begin{bmatrix} u(x_1,y_1) & u(x_1,y_2) \\ u(x_2,y_1) & u(x_2,y_2) \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} = \begin{bmatrix} c_1 + d_1 & c_1 - d_1 \\ c_2 - d_2 & c_2 + d_2 \end{bmatrix}$$

It is called the *confusion matrix* in the context of hypothesis testing, where rows correspond to the true states $\{x_1, x_2\}$, and columns correspond to accepting or rejecting the hypothesis $\{y_1, y_2\}$. The set of all joint distributions $W(x,y)$ is a 3-simplex (tetrahedron), shown in Figure 1. The 2D surface in the middle is the set of all product distributions $W(x,y;0) = P(x)Q(y)$, which correspond to the minimum $I(X,Y) = 0$ of mutual information (independent x,y). With no additional information about x, the decision y_1 to accept or y_2 to reject the hypothesis is completely determined by the utilities and prior probabilities $P(x_1) = p$ and $P(x_2) = 1 - p$. Thus, one has to compare expected utilities $\mathbb{E}_P\{u \mid y_1\} = p\,u_{11} + (1-p)\,u_{21}$ and $\mathbb{E}_P\{u \mid y_2\} = p\,u_{12} + (1-p)\,u_{22}$. The output distribution $Q(y)$ is an elementary δ-distribution:

$$Q(y_1) = \begin{cases} 1 & \text{if } \frac{p}{1-p} \geq \frac{u_{22}-u_{21}}{u_{11}-u_{12}} = \frac{d_2}{d_1} \\ 0 & \text{otherwise} \end{cases}$$

The optimal value corresponding to $I = 0$ information is $U(0) = p\,c_1 + (1-p)\,c_2 + |p\,d_1 - (1-p)\,d_2|$. In the case when $c_1 = c_2 = c$ and $d_1 = d_2 = d$, the condition for y_1 is $d(2p-1) \geq 0$ and $U(0) = c + d|2p-1|$. With $c = 1/2$ and $d = 1/2$, the value $U(0) = \frac{1}{2} + \frac{1}{2}|2p-1|$ represents the best possible *accuracy* for prior probabilities $P(x) \in \{p, 1-p\}$. If additional information about x is communicated, say by some random variable $z \in Z$, then the maximum possible improvement $V(I) = U(I) - U(0)$ is the value of this information. The first step in deriving function $U(I)$ for the value of Shannon's information (2) is to obtain the expression for function $\Gamma(\beta) = \Gamma_0(\beta) - H(X)$.

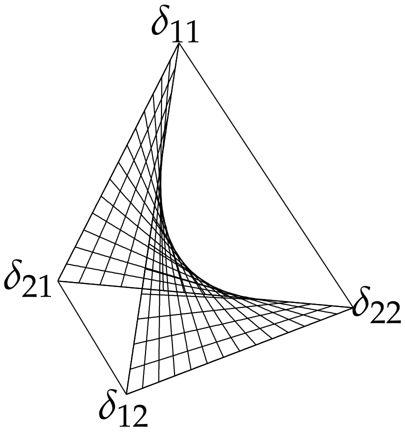

Figure 1. A 3-simplex of all joint distributions on a 2 × 2 system.

Writing Equation (4) in the matrix form $\|e^{\beta\,u(x,y)}\|^T P(x)\,e^{-\gamma(\beta,x)} = 1$ and using the inverse matrix $(\|e^{\beta\,u(x,y)}\|^T)^{-1}$ gives the solution for function $e^{-\gamma_0(\beta,x)} = P(x)e^{-\gamma(\beta,x)}$:

$$\begin{bmatrix} p\,e^{-\gamma(\beta,x_1)} \\ (1-p)\,e^{-\gamma(\beta,x_2)} \end{bmatrix} = \begin{bmatrix} e^{\beta\,u_{11}} & e^{\beta\,u_{21}} \\ e^{\beta\,u_{12}} & e^{\beta\,u_{22}} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\det\|e^{\beta\,u}\|^T} \begin{bmatrix} e^{\beta\,u_{22}} & -e^{\beta\,u_{21}} \\ -e^{\beta\,u_{12}} & e^{\beta\,u_{11}} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where $\det\|e^{\beta\,u}\|^T = e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})} = 2e^{\beta\,(c_1+c_2)}\sinh[\beta\,(d_1+d_2)]$. This gives two equations:

$$p\,e^{-\gamma(\beta,x_1)} = \frac{e^{\beta\,u_{22}} - e^{\beta\,u_{21}}}{e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})}} = e^{-\beta\,c_1}\,\frac{\sinh(\beta\,d_2)}{\sinh[\beta\,(d_1+d_2)]} =: e^{-\gamma_0(\beta,x_1)}$$

$$(1-p)\,e^{-\gamma(\beta,x_2)} = \frac{e^{\beta\,u_{11}} - e^{\beta\,u_{12}}}{e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})}} = e^{-\beta\,c_2}\,\frac{\sinh(\beta\,d_1)}{\sinh[\beta\,(d_1+d_2)]} =: e^{-\gamma_0(\beta,x_2)}$$

Therefore, the expression for function $\Gamma_0(\beta) := p\gamma_0(\beta,x_1) + (1-p)\gamma_0(\beta,x_2)$ is

$$\Gamma_0(\beta) = \beta\,[p\,c_1 + (1-p)\,c_2] + \ln|\sinh[\beta\,(d_1+d_2)]| - p\,\ln|\sinh(\beta\,d_2)| - (1-p)\,\ln|\sinh(\beta\,d_1)|\,.$$

Its first derivative $\Gamma_0'(\beta)$ gives the expression for $U(\beta)$:

$$U(\beta) = p\,c_1 + (1-p)\,c_2 + (d_1+d_2)\coth[\beta\,(d_1+d_2)] - p\,d_2\coth(\beta\,d_2) - (1-p)\,d_1\coth(\beta\,d_1)\,.$$

The expression for information is obtained from $I(\beta) = H(X) - [\Gamma_0(\beta) - \beta\,\Gamma_0'(\beta)]$, where $H(X) = -p\,\ln p - (1-p)\,\ln(1-p)$. Two functions $U(\beta)$ and $I(\beta)$ define parametric dependency $U(I)$ for the value of Shannon's information (2).

Notice that function $\Gamma_0(\beta)$ (and hence $U(\beta)$ and $I(\beta)$) depends in general on $P(x) \in \{p, 1-p\}$. If, however, $c_1 = c_2 = c$ and $d_1 = d_2 = d$, then, using the formula $\frac{\sinh(2x)}{\sinh(x)} = 2\cosh(x)$, we obtain simplified expressions: $\Gamma_0(\beta) = \beta\,c + \ln[2\cosh(\beta\,d)]$ and

$$U(\beta) = c + d\,\tanh(\beta\,d)\,, \qquad I(\beta) = H(X) - [\ln[2\cosh(\beta\,d)] - \beta\,d\,\tanh(\beta\,d)]\,.$$

Let us denote $\theta := \frac{U-c}{d} = \tanh(\beta\,d) \in [0,1]$. Then the expression for information is

$$\begin{aligned} I(\theta) &= H(X) - \ln[2\cosh(\tanh^{-1}\theta)] + \theta\,\tanh^{-1}\theta \\ &= H(X) + \ln\frac{1}{2} + \frac{1}{2}\ln(1-\theta^2) + \frac{1}{2}\theta\,\ln\frac{1+\theta}{1-\theta} \\ &= H_2[p] - H_2\left[\frac{1+\theta}{2}\right]. \end{aligned}$$

In the first step we used the formulae $\cosh(\tanh^{-1}\theta) = \frac{1}{\sqrt{1-\theta^2}}$ and $\tanh^{-1}\theta = \frac{1}{2}\ln\frac{1+\theta}{1-\theta}$. The last equation is written using binary entropies $H_2[p] = -p\,\ln p - (1-p)\,\ln(1-p)$, which shows that an increase of information in a binary system is directly related to an increase of the probability $(1+\theta)/2 \geq \max\{p, 1-p\}$ due to conditioning on the 'message' $z \in Z$ about the realization of $x \in X$. Additionally, substituting $\theta = (U-c)/d$ we obtain the closed-form expression:

$$I(U) = H_2[p] - H_2\left[\frac{1}{2} + \frac{1}{2}\frac{U-c}{d}\right] \tag{6}$$

Let us derive the equations for the output probabilities $Q(y) = \sum_x P(y \mid x)\,P(x)$. This can be done using Equation (5), which in the matrix form is $\|e^{\beta\,u(x,y)}\|Q(y) = e^{\gamma(\beta,x)}$. Thus, we obtain

$$\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} e^{\beta\,u_{11}} & e^{\beta\,u_{12}} \\ e^{\beta\,u_{21}} & e^{\beta\,u_{22}} \end{bmatrix}^{-1} \begin{bmatrix} e^{\gamma(\beta,x_1)} \\ e^{\gamma(\beta,x_2)} \end{bmatrix} = \frac{1}{\det\|e^{\beta\,u}\|} \begin{bmatrix} e^{\beta\,u_{22}} & -e^{\beta\,u_{12}} \\ -e^{\beta\,u_{21}} & e^{\beta\,u_{11}} \end{bmatrix} \begin{bmatrix} e^{\gamma(\beta,x_1)} \\ e^{\gamma(\beta,x_2)} \end{bmatrix},$$

where $\det\|e^{\beta\,u}\| = e^{\beta\,(u_{11}+u_{22})} - e^{\beta\,(u_{12}+u_{21})} = 2e^{\beta\,(c_1+c_2)}\sinh[\beta\,(d_1+d_2)]$. This gives two equations:

$$Q(y_1) = \frac{p}{1-e^{-2\beta\,d_2}} + \frac{1-p}{1-e^{2\beta\,d_1}}\,, \qquad Q(y_2) = \frac{1-p}{1-e^{-2\beta\,d_1}} + \frac{p}{1-e^{2\beta\,d_2}}\,.$$

It is easy to check that $Q(y_1) + Q(y_2) = 1$. Additionally, if $p = 1-p$, then $Q(y_1) \geq 0$ and $Q(y_2) \geq 0$ for all $\beta \geq 0$. However, when $p \neq 1-p$, there exists $\beta_0 > 0$ such that either $Q(y_1) < 0$ or $Q(y_2) < 0$ for $\beta \in [0, \beta_0)$. The value β_0 can be found from $Q(y_1) = 0$ or $Q(y_2) = 0$. For $d_1 = d_2 = d$ this value is

$$\beta_0 = \frac{1}{2d}\left|\ln\left(\frac{p}{1-p}\right)\right|.$$

One can show that $I(\beta_0) = 0$ and $U(\beta_0) = c + d|2p-1|$. Thus, the output probabilities are non-negative for all $\beta \geq \beta_0$, which corresponds to positive information $I \geq 0$ and $U(I) \geq U(0)$.

It is important to note that in the limit $\beta \to \infty$, corresponding to an increase of information to its maximum, the output probabilities $Q(y) \in \{q, 1-q\}$ converge to $P(x) \in \{p, 1-p\}$.

3. Application: Accuracy of Time-Series Forecasts

In this section, we illustrate how the value of information can facilitate the analysis of the performance of data-driven models. Here we use financial time-series data and predict the signs of future log returns. Thus, if $s(t)$ and $s(t-1)$ are prices of an asset at two time moments, then $r(t) = \ln[s(t)/s(t-1)]$ is the log-return at t. The models will try to predict whether the future log return $r(t+1)$ is positive or negative. Thus, we have a 2 × 2 system, where $x \in \{x_1, x_2\}$ is the true sign, and $y \in \{y_1, y_2\}$ is the prediction. The accuracy of different models will be evaluated against the theoretical upper bound, defined by the value of information.

The data used here are from the set of close-day prices $s(t)$ of several cryptocurrency pairs between 1 January 2019 and 11 January 2021. Figure 2 shows the price of Bitcoin against USD (left) and the corresponding log returns (right). Predicting price changes is very challenging. In fact, in economics, log returns are often assumed to be independent (and hence prices $s(t)$ are assumed to be Markov). Indeed, one can see no obvious relation on the left chart on Figure 3, which plots logreturns $r(t)$ (abscissa) and $r(t+1)$ (ordinates). In reality, however, some amounts of information and correlations exist, which can be seen from the plot of the autocorrelation function for BTC/USD shown on the right chart of Figure 3.

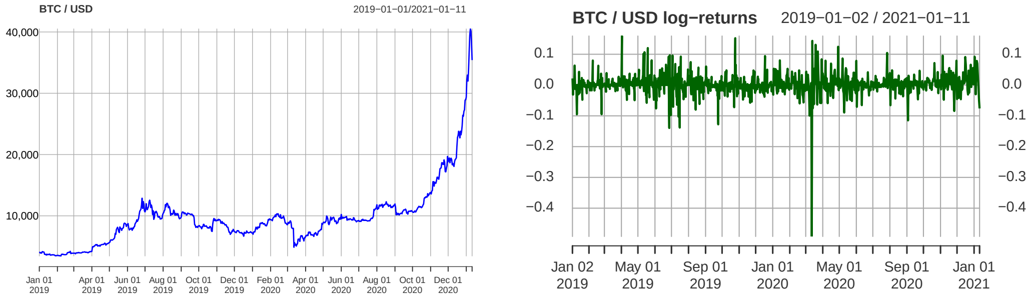

Figure 2. Close day prices of BTC/USD (**left**) and the corresponding log returns (**right**).

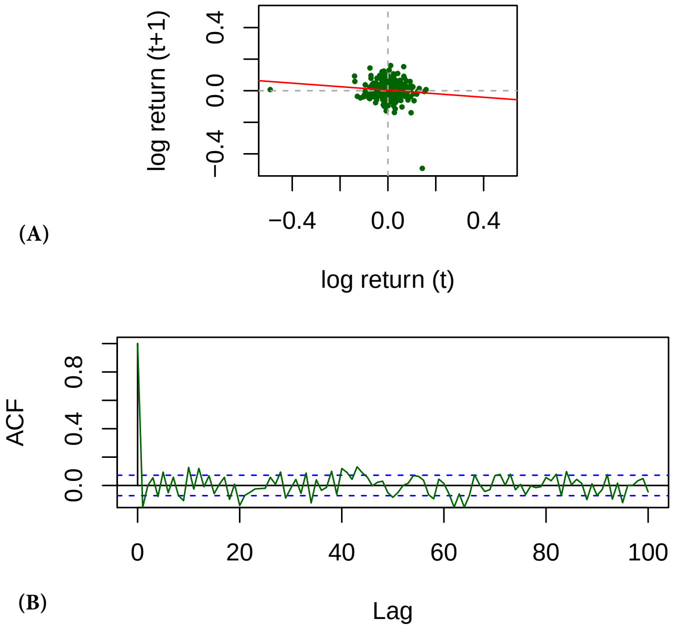

Figure 3. Log returns of BTC/USD on two consecutive days (**A**); the autocorrelation function (**B**).

The idea of autoregressive models is to use the small amounts of information between the past and future values for forecasts. In addition to autocorrelations (correlations between the values of $\{r(t)\}$ at different times), information can be increased by using cross-correlations (correlations between log-returns of different symbols in the dataset). Thus, the vector of predictors used here is an $m \times n$-tuple, where m is the number of symbols used, and n is the number of time lags. In this paper, we report the results of models using the range $m \in \{1, 2, \ldots, 5\}$ of symbols (BTC/USD, ETH/USD, DAI/BTC, XRP/BTC, IOT/BTC) and $n \in \{2, 3, \ldots, 20\}$ of lags. This means that the models used predictors $(z_1, \ldots, z_{m\times n})$, where $m \times n$ ranged from 2 to 100. The model output $y(z)$ is the forecast of the sign $x \in \{-1, 1\}$ (the response) of future log return $r(t+1)$ of BTC/USD. Here we report results from the following models:

1. Logistic regression (LM). This model has no hyperparameters.
2. Partial least squares discrimination (PLSD). We used the SIMPLS algorithm [7] with three components.
3. Feed-forward neural network (NN). Here we used one hidden layer with three logistic units.

In order to analyse the performance of models using the value of information, one has to estimate the amount of information between the predictors $z_1, \ldots, z_{m\times n}$ and the response variable x. Here we employ two methods. The first uses the following Gaussian formula [4]:

$$I(X, Z) \approx \frac{1}{2}[\ln\det K_z + \ln\det K_x - \ln\det K_{z\oplus x}]\,,$$

where K_i are the covariance matrices. Because the distributions of log returns are generally not Gaussian, this formula is an approximation (in fact, it gives a lower bound). The second method is based on the discretization of continuous variables. Because models were used to predict signs of log returns, here we used discretization into two subsets. Figure 4 shows the average amounts of information $I(X, Z)$ in the training sets, computed using the Gaussian formula (left) and using binary discretization (right). Information (ordinates) is plotted against the number n of lags (abscissa) and for $m \in \{1, 2, \ldots, 5\}$ symbols (different curves). One can see that the amounts of information using Gaussian approximation (left) are generally lower than those using discretization (right). We note, however, that linear models can only use linear dependencies (correlations), which means that Gaussian approximation is sufficient for assessing the performance of linear models, such as LM and PLSD. Non-linear models, on the other hand, can potentially use all information present in the data. Therefore, we used information estimated with the second method to assess the performance of NN.

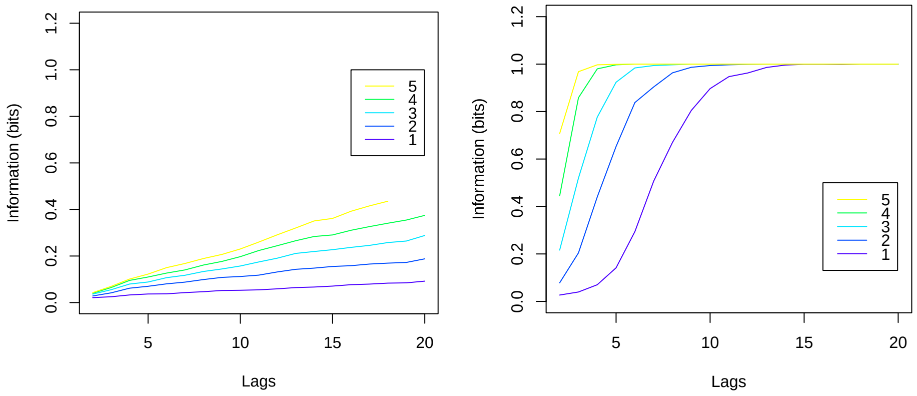

Figure 4. The average amounts of mutual information between predictors and response in the training sets, computed using Gaussian approximation (**left**) and using binary discretization (**right**). The abscissa shows the numbers n of lags; different curves correspond to numbers m of symbols used.

For each collection of predictors $(z_1, \ldots, z_{m\times n})$ and response x, the data were split into multiple training and testing subsets using the following rolling window procedure: we used 200- and 50-day data windows for training and testing, respectively; after training and testing the models, the windows were moved forward by 50 days and the process repeated. Thus, the data of approximately 700 days (January 2019 to January 2021) were split into $(700 - 200)/50 = 10$ pairs of training and testing sets. The results reported here are the average of results from these 10 subsets.

Figure 5 shows the accuracies of models plotted against information amounts I in the training data. The top row shows results on the training sets (i.e. fitted values) and the bottom row for new data (i.e., predicted values). Different curves are plotted for different numbers of symbols $m \in \{1, \dots, 5\}$. The theoretical upper bounds are shown by the Accuracy(I) curves computed using the inverse of function (6) with $c = d = 1/2$ and $p = 1/2$. Here we note the following observations:

1. The accuracy of fitting the training data closely follows theoretical curve Accuracy(I). The accuracy of predicting new data (testing sets) is significantly lower.
2. Increasing information increases the accuracy on training data, but not necessarily on new data.
3. Models using $m > 1$ symbols appear to achieve better accuracy than models using $m = 1$ symbol with the same amounts of information. Thus, surprisingly, cross-correlations potentially provide more valuable information for forecasts than autocorrelations.

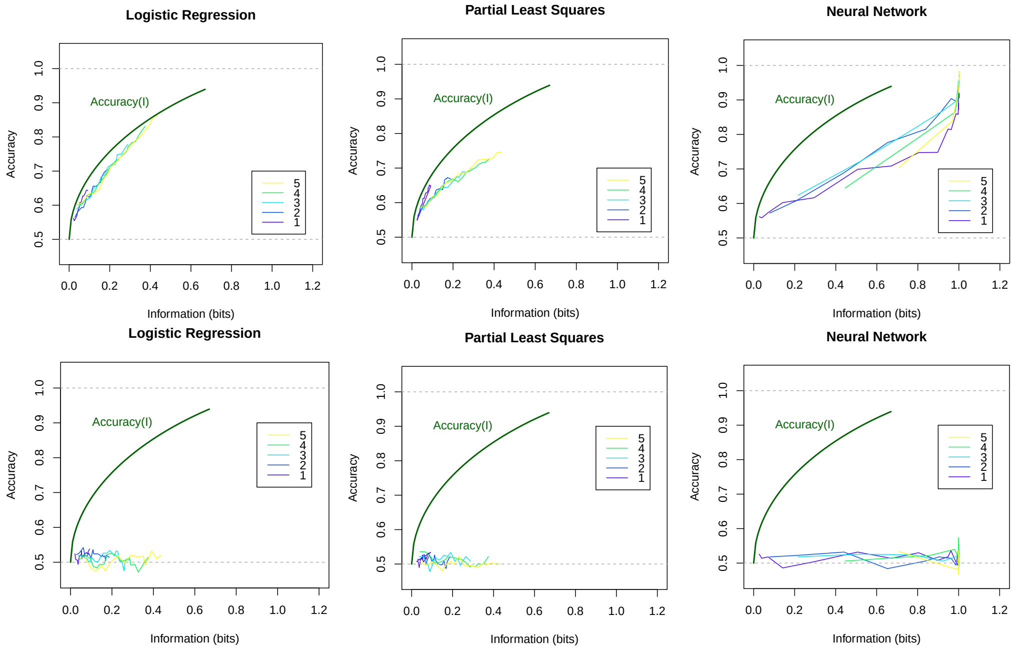

Figure 5. Accuracy of fitted values on training data (top row) and of predicted values on testing data (bottom row) for three types of models plotted as functions of information in the training data. Theoretical curves are plotted using the inverse of function (6) for $c = d = 1/2$ and $p = 1/2$. Different curves correspond to the number m of symbols used.

4. Discussion

We have reviewed the main ideas of Stratonovich's value of information theory [2,4] and applied it to the simplest 2×2 Bayesian system. We explicitly performed the main computations for the cumulant generating function $\Gamma(\beta) = \Gamma_0(\beta) - H(X)$ and derived functions $U(\beta)$ and $I(\beta)$ defining the dependency $U(I)$ and the value of Shannon's information $V(I) = U(I) - U(0)$. The main application of the considered binary example the is evaluation of the accuracy of model predictions or hypothesis testing. The analysis the of performance of data-driven models can be enriched by the use of the value of information. However, one needs to be careful about the estimation of the amount of information in the data. Gaussian approximation of mutual information can be used for linear models. However, other techniques should be used for the analysis of non-linear models, such as neural networks. Here we applied the value of information to the analysis of financial time-series forecasts. These methods can be generalized to many other machine learning and data science problems.

Author Contributions: Conceptualization and formal analysis, R.B.; methodology, R.B., P.P. and J.P.; software, R.B.; modelling and experiments, R.B.; data curation, R.B., P.P. and J.P.; writing—original draft preparation, R.B.; writing—review and editing, R.B., P.P. and J.P.; funding acquisition, J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the ONR grant number N00014-21-1-2295.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Close day prices used in this work are available at https://doi.org/10.22023/mdx.21436248.

Acknowledgments: Stefan Behringer is deeply acknowledged for additional discussion of the example, and Roman Tarabrin is deeply acknowledged for providing a MacBookPro laptop used for the computational experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Howard, R.A. Information Value Theory. *IEEE Trans. Syst. Sci. Cybern.* **1966**, *2*, 22–26. [CrossRef]
2. Stratonovich, R.L. On value of information. *Izv. USSR Acad. Sci. Tech. Cybern.* **1965**, *5*, 3–12. (In Russian)
3. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423 and 623–656. [CrossRef]
4. Stratonovich, R.L. *Theory of Information and Its Value*; Springer: Cham, Switzerland, 2020.
5. Belavkin, R.V. Relation Between the Kantorovich-Wasserstein Metric and the Kullback-Leibler Divergence. In Proceedings of the Information Geometry and Its Applications, Liblice, Czech Republic, 12–17 June 2016; Ay, N., Gibilisco, P., Matúš, F., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 363–373.
6. Belavkin, R.V. Optimal measures and Markov transition kernels. *J. Glob. Optim.* **2013**, *55*, 387–416. [CrossRef]
7. de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263. [CrossRef]